Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 100 (2016) 686 – 692

Procedia Computer Science

Conference on ENTERprise Information Systems / International Conference on Project MANagement / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2016, October 5-7, 2016

Towards an infodemiological algorithm for classification of Filipino health tweets

Kennedy Espina[a] *, Ma. Regina Justina Estuar[a], Delfin Jay Sabido IX[b], Raymond Josef Edward Lara[b], Vikki Carr de los Reyes[c]

[a]*Ateneo de Manila University, Katipunan Avenue, Quezon City, 1108, Philippines*
[b]*IBM Philippines R&D Laboratory, Eastwood City Cyberpark, Quezon City, 1110, Philippines*
[c]*Department of Health, San Lazaro Compound, Manila City, 1014, Philippines*

Abstract

Finding innovative ICT solutions to enhance the Philippines' health sector is part and parcel of the *Philippine eHealth Strategic Framework and Plan 2020 program*. This study sees the opportunity of using collected Twitter data to create a model that processes tweets to produce a dataset that may be relevant in the field of epidemiology and *infodemiology*. Through the collection of relevant tweets, future studies may make use of the output of this research for various purposes, such as the improvement of epidemiological systems of the Department of Health in support of the eHealth strategy. In this study, we used the Naïve-Bayes classification model, an efficient text classifier, to create a model that determines whether a tweet is "infodemiological" or not. From the collected 18,044 tweets, we have narrowed it down to 1,090 tweets (6.04%) that can be used in epidemiology. Using this as a dataset for training and testing, the model was able to classify 79.91% of tweets correctly. This research shows that it is indeed feasible to collect and classify enough infodemiological tweets in the Filipino language, which in turn can be used for future infodemiological studies.

Peer-review under responsibility of the organizing committee of CENTERIS 2016

Keywords: Epidemiology; Infodemiology; Philippines; Tweets; Modeling; Classification

* Corresponding author. Tel.: +63-998-9701831.
E-mail address: kennedyespina@yahoo.com

Peer-review under responsibility of the organizing committee of CENTERIS 2016
doi:10.1016/j.procs.2016.09.212

1. Introduction

In the era of online social networking, it can be said that people are more expressive than ever since there is now an avenue to publicly express emotions (sentiments and reactions) and actions (behavior, location) in real-time[1]. Through having these dimensions, social networking websites such as Twitter have become prime candidates to learn more about the daily conditions of people, which can be used for marketing, disaster controls, and tracking health emergencies[2,3]. Twitter has been used in different parts of the world to track different diseases. The word "infodemiology," a portmanteau of the word "information" and "epidemiology," comes into mind when using online posts for public health. Infodemiology deals with people's online "behavior" when posting disease-related information, more than the actual tracking of diseases themselves.

In the context of the Philippines, a country considered to be the social networking capital of the world, we see the potential of Twitter posts as an added layer to improve on existing epidemiological systems used in the country. This research is aligned with the *Philippine eHealth Strategic Framework and Plan 2020* that may help in strengthening existing eHealth solutions through adding to the *Information Sources* pillar of the framework. Through the use of ICT, the state of eHealth in the Philippines should be improved to provide more efficient universal healthcare to people. The use of Twitter, in this respect, with the proper algorithm, may give the Department of Health a way to improve on their existing disease surveillance systems, since tweets are collected real-time.

The common way of tracking diseases, particularly communicable diseases, requires work that takes up a lot of time, which in turn would lead to the delay in analysis and reporting of results[4]. In the Philippines, there are several types of systems activated by the national Department of Health depending on the event or the need that arises. For example, the *Surveillance in Post Extreme Emergencies and Disasters (SPEED)* is activated when disasters strike an area of the Philippines. This helps in creating a population-based surveillance on the conditions of people in evacuation areas. Although these systems are efficient, all these take time before the data actually reaches and get processed by the Department of Health. With the use of the model created in this research, data can be processed proactively rather than reactively. The model could then be used by entities such as the Department of Health to complement and improve their existing systems.

Since we are using tweets, the concept of Big Data comes into play, and for this research, the focus is on the volume and veracity of data. The research's objectives are 1) to show that the *volume* of tweets collected is sufficient and enough to be used for infodemiological studies in the Filipino language, and 2) to show that it is feasible to create a classifier to aid in tagging whether a tweet is infodemiological in nature or not, considering the *veracity* of the data collected.

2. Review of related literature

2.1. Current Philippine eHealth systems

eHealth, as defined by the World Health Organization (WHO), is the utilization of Information and Communications Technology (ICT) solutions in the field of healthcare. The use of ICT in the Philippines has been given priority in the *National Objectives for Health* by the Department of Health since the year 2005. In the year 2014, the *Philippine eHealth Strategic Framework and Plan 2020* (Referred to as *eHealth 2020* hereafter) was introduced[5].

As seen in Figure 1, there are different aspects of public health the eHealth 2020 framework is trying to address. For this research, we are focusing on *eHealth solutions*, particularly that of Information Sources. We see tweets as a way to collect information rapidly for syndromic surveillance. Currently there are two (2) systems where an integration and harmonization of the output may be put into use. These are *Electronic Field Health Services and Information System (eFHSIS)* and the *Event-Based Surveillance and Response (ESR)*. The eFHSIS is designed to record patient data at the *barangay* – or village - level. This system is made for the detection of diseases through syndromic surveillance, since the data captured includes general consultation records, clinical diagnosis, immunizations, and more. On the other hand, the ESR system is used to capture all types of health events with potential public health threat (including outbreaks) for appropriate response. It uses six core processes of *capture*, *filter*, *verify*, *assessment*, *response* and *feedback*, to confirm whether the reports, both formal and informal, are correct.

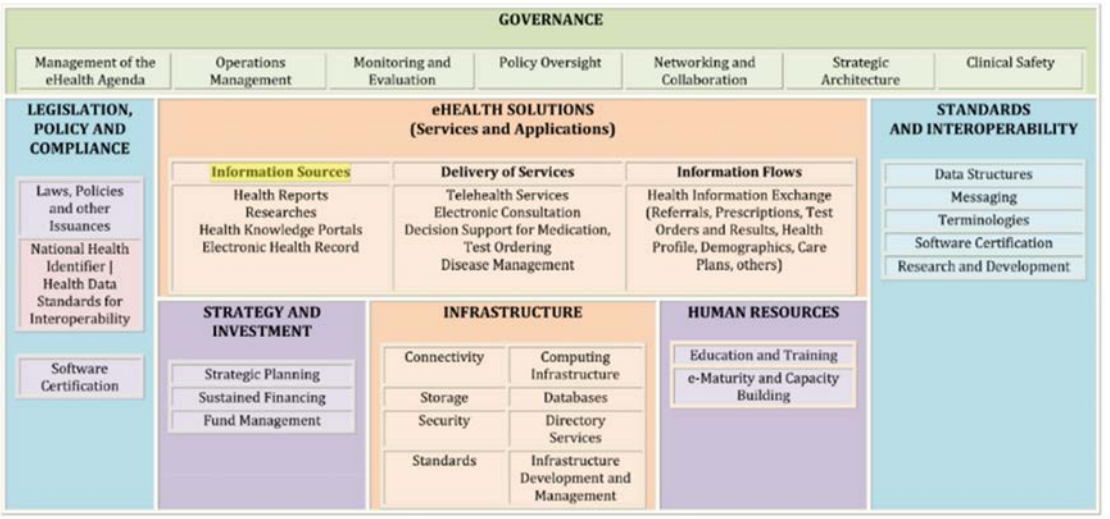

Fig. 1. Philippine eHealth Strategic Framework and Plan 2020[5].

2.2. Introducing infodemiology

Infodemiology, a portmanteau of "Information" and "Epidemiology", aims to measure the pulse of people's public opinions, behavior, and knowledge through tracking their online activities[6]. In this sense, it creates an "epidemiological" analysis of the information being created by users online. The main advantage of using infodemiological methodologies is the relatively fast and easy collection of data[6]. Added to this, the dataset collected using infodemiology may contain data that have not been gathered through using traditional methodologies due to the wider search space, which is really the Internet. Sources for Infodemiology may come in the form of web queries[7], news articles, and social networking websites, such as Twitter.

With misinformation being widely sent on the Internet[8], the use of "*infoveillance,*" or information surveillance, is needed to countercheck data's validity as a possible source of health data to be used in researches. An infodemiological research focusing on Flu-related tweets shows how misinformation can be minimized by correlating results with official health records[9]. The research was done in Portugal and showed the feasibility of using such system in estimating and predicting the incidence rate of influenza-like illness in the country[9].

2.3. Feasibility of tweet mining in health

Some researches use already existing systems to aid on the collection and visualization of tweets. One example used is a system developed by M-Eco, or Medical Ecosystem[10]. The system used by M-Eco is responsible for the

collection of data from social media websites until the visualization process of the results. Algorithms can then be created to detect "signals" or "anomalies" from the gathered data through clustering. For this research, the same methodology can be utilized through a tool developed using the Twitter API to collect real-time tweets. Aside from this, the research would apply various text mining methodologies to ensure that the output of the algorithm is valid.

2.4. Using tweets to track HIV and Dengue

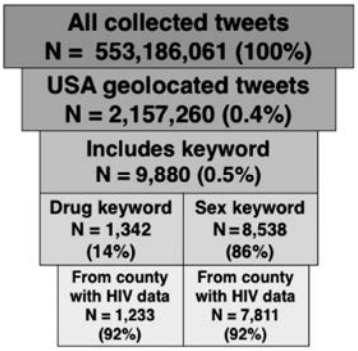

Fig. 2. Twitter Data Filtering by Young et al[11].

The spread of the Human Immunodeficiency Virus (HIV) was tracked using tweets by collecting posts that suggest sexual or drug-related activities[11]. The collected tweets were passed through different phases of filtering to get the desired output. The last level of filtering shows that they only used geo-located tweets that are found in US counties that have official health records found in *aidsvu.org*. These records are used as the gold standard to verify the collected tweets using univariate regression. This research follows the framework found in Fig 2. In another study, a framework was developed focusing on tracking the Dengue fever in Brazil[4]. The research used four (4) parameters in the filtering of the collected tweets. These are *volume*, *location*, *time*, and *public perception*[4].

3. Methodology

3.1. Tweet extraction

The tool developed using the Twitter API is utilized to collect tweets as they become available. The tool allows the capturing of tweets in real-time based on keywords specified in its configuration file. The keywords used, consisting of both Filipino and English words, are based on the framework by Young, which focuses on certain symptoms usually found in different diseases. The keywords are then subdivided into three (3) categories, namely *Colloquial Terms*, *Symptoms*, and *Behavior/Action*[10]. Examples can be seen in Table 1.

Table 1. Categories of keywords used to search tweets

Category	Description	Examples
Coloquial Terms	Terms that indicate the presence of a disease.	*"Lagnat (Flu/Sick)", "Flu", "Sick", "Feeling Warm"*
Symptoms	Symptoms of diseases that can be used collectively for syndromic surveillance.	*"Ubo (Cough)", "Coughing", "Sipon (Colds)", "Sneezing"*
Behavior/Actions	Behavior of people when they feel sick. Example: buying medicines, filing sick leaves	*"Mercury Drug (A drugstore in the Philippines)", "Sick Leave Na Naman ("Filing a sick leave again")"*

3.2. Data tagging and cleansing

The R Programming Language and its packages were used to process the collected data from Twitter because of its extensive package for text mining. After collecting tweets using the keywords, we filtered the tweets so that we only have those that are language-tagged as *"tl"* for Tagalog/Filipino. The tweets in JSON format are then imported to R and are converted into a Data Frame Object. From the dataset, texts with *"rt"* or those that are indicated as retweeted, are removed. Using an exported CSV file, the tweets are manually tagged whether it is "infodemiological" in nature (Tagged **"TRUE"**) or not (Tagged **"FALSE"**). The updated CSV file is then re-imported to R as a Data Frame object ready for processing. A corpus was made using the tweets in the Data Frame. Using the *tm* package for R, data cleansing was done to the created corpus, such as the removal of numbers, lowering the case of texts, and removing of stopwords. R already has a library for stopwords, but since we're dealing with bilingual text, we added other Filipino stopwords as well.

4. Results and discussions

4.1. Manual tagging

Such as the method used by another study[11], several layers for filtering was used to arrive at the final set of tweets tagged as *"true"* or *"false,"* depending on whether they are infodemiological or not, respectively. Table 2 shows a breakdown of tweets from the original 18,044 tweets up to the 1,090 tweets are tagged as infodemiological.

Table 2. Breakdown of Tweets

Category	Number of Tweets
Collected Tweets	18,044
Filipino-Tagged Tweets	4,293
After Removal of Retweets	2,788
Manually Tagged as Infodemiological in Nature	1,090

Tagging is dependent on two categories: 1) Physical manifestations of symptoms or 2) Relevant actions when getting sick. Table 3 shows samples of tweets from these categories.

Table 3. Sample Tweets that are tagged true

Category	Sample Tweets
Physical Manifestation of Symptoms	*"Sakit ng ulo ko"* (Translation: "*My head hurts"*)
	"Ubo + Sipon = ???" (Translation: *"Cough + Sneezing = ???"*)
Relevant actions when getting sick	*"…tuwing sinisipon ako, laging binibili ng lola ko neozep"* (Translation: *"… everytime I get the colds, my grandmother always buys neozep"*) Note: **Neozep** is a medicine for treating colds

From the data collected after manual tagging, there are lower amount of tweets (1,090 tweets in one day) that may be used for infodemiological studies compared to the number Gomide[4] used in their research (Average of 3,054 tweets per day)[2]. While this is true, it should be noted that these 1,090 tweets are the ones already tagged whether they are disease-related or not. In the study, the daily 3,054 were every tweet they collected with the word "Dengue" in them.

4.2. Naïve-Bayes classification

The Naïve-Bayes classification is a probabilistic model that is proven to be efficient in text classification[12]. This was the chosen algorithm as it is able to process training datasets faster and more efficiently compared to other classification algorithm[13] for the purpose of this research. Using the E1071 package for R, a Naïve-Bayes classification algorithm was used to train and test the data to determine whether a tweet is infodemiological in nature or not. Seen in Table 4 is the result of the simulation in R.

Table 4. Cross-table of Results after classification

	Actual Infodemiological	**Actual Non-Infodemiological**	**Total Predicted Tweets**
Predicted Infodemiological	**215 Tweets** Precision: 71.67%	**55 Tweets**	**270 Tweets**
Predicted Non-Infodemiological	**85 Tweets**	**342 Tweets** Precision: 86.15%	**427 Tweets**
Total Actual Tweets	**300 Tweets**	**397 Tweets**	**697 Tweets**

For the Naïve-Bayes classification, 75% (2,091 tweets) of the manually tagged tweets were used as the training dataset, and the remaining 25% (697) were used for testing. There is an unequal amount of tweets tagged as "TRUE" (1699 tweets) and tagged as "FALSE" (1,089 tweets). The tweets are randomly distributed for both the training and test datasets.

As seen in Table 3, the algorithm's precision in predicting a "FALSE" is at **86.15%** *(342 tweets/397 tweets),* but the prediction overestimates by **107%** *(427 tweets/397 tweets).* On the other hand, the algorithm's precision in predicting a "TRUE" is at **71.67%** (215 tweets/300 tweets) and it underestimates at **90%** (270 tweets/300 tweets). The nuances in the Filipino language may have contributed to the result of the algorithm. Overall, the algorithm was able to correctly classify infodemiological and non-infodemiological tweets **79.91%** (557 out of the 697 tweets) of the time.

Through using the Naïve-Bayes classification algorithm, it was shown that it is feasible to create an effective classifier to aid in identifying tweets that are epidemiological in nature.

5. Conclusion and recommendation for future studies

This research aims to contribute to the *Information Sources* pillar of the eHealth 2020 vision of the Philippine government. Through utilizing Twitter, along with the advantages it brings such as real-time collection, it is presented that it is feasible to collect enough tweets posted in the Filipino language that are infodemiological in nature. This kind of researches had already been done in different parts of the world, and it is about time that the Philippines utilizes this tool as well. Added to this, we have shown that through using a Naïve-Bayes classification algorithm, we can create a classifier to identify tweets are relevant or not to infodemiology.

After showing how there is potential in the volume of tweets collected, we recommend having further research into improving the algorithm for classification used in this research. Added to this, we highly recommend doing research to see possible usage of processing tweets in a larger scale in the Philippines, particularly in the field of epidemiology.

Acknowledgements

We would like to primarily thank the Philippines' Department of Science and Technology - Philippine Council for Health Research and Development for extending their resources for this research. We would also like to thank the Philippines' Department of Health, IBM Philippines R&D Laboratory, Ateneo de Manila University – Department of

Information Systems and Computer Science, Ateneo Java Wireless Competency Center, Ateneo Social Computing Sciences Laboratory, and Mr. John Noel Victorino for lending their knowledge for this research.

References

1. Deller R. Twittering on: Audience research and participation using Twitter. Participations. 2011 May;8(1):216-45.
2. Zeng D, Chen H, Lusch R, Li SH. Social media analytics and intelligence. Intelligent Systems, IEEE. 2010 Nov;25(6):13-6.
3. Ribarsky W, Wang DX, Dou W. Social media analytics for competitive advantage. Computers & Graphics. 2014 Feb 28;38:328-31.
4. Gomide J, Veloso A, Meira Jr W, Almeida V, Benevenuto F, Ferraz F, Teixeira M. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. InProceedings of the 3rd international web science conference 2011 Jun 15 (p. 3). ACM.
5. Philippines ehealth strategic framework and plan 2014-2020. Online, 2013.
6. Eysenbach G. Infodemiology and infoveillance: tracking online health information and cyberbehavior for public health. American journal of preventive medicine. 2011 May 31;40(5):S154-8.
7. Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. InAMIA Annual Symposium Proceedings 2006 (Vol. 2006, p. 244). American Medical Informatics Association.
8. Eysenbach G. Infodemiology: the epidemiology of (mis) information. The American journal of medicine. 2002 Dec 15;113(9):763-5.
9. Santos JC, Matos S. Analysing Twitter and web queries for flu trend prediction. Theoretical Biology and Medical Modelling. 2014 May 7;11(Suppl 1):S6..
10. Denecke K, Krieck M, Otrusina L, Smrz P, Dolog P, Nejdl W, Velasco E. How to exploit twitter for public health monitoring. Methods Inf Med. 2013 Jan 1;52(4):326-9.
11. Young SD, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. Preventive medicine. 2014 Jun 30;63:112-5.
12. Rennie JD, Shih L, Teevan J, Karger DR. Tackling the poor assumptions of naive bayes text classifiers. InICML 2003 Aug 22 (Vol. 3, pp. 616-623).
13. Narayanan V, Arora I, Bhatia A. Fast and accurate sentiment classification using an enhanced Naive Bayes model. InIntelligent Data Engineering and Automated Learning–IDEAL 2013 2013 Oct 20 (pp. 194-201). Springer Berlin Heidelberg.